ARC File Format Revision 3.0: Feedback from the Los Alamos National Laboratory

Xiaoming Liu <liu_x@lanl.gov>
Herbert Van de Sompel <herbertv@lanl.gov>

1. Introduction

This document provides feedback to the document:

> Title: ARC file Revision 3.0 Proposal
> Authors: Steen Christensen, Michael Stack
> Editor: Michael Stack
> Date: 09/09/2004
> Revision: 1

Although the motivation for providing feedback to the proposed ARC file format revision is related to the desire to use the ARC file format for the storage of "local" (not crawled) content in the Repository of the LANL Research Library, most comments provided here are of a general nature. As a matter of fact, all requirements as listed in Section 1 of the aforementioned document have been taken into account for the feedback provided here.

In addition to that, applicability of the ARC file format for a "local" use case, such as the one at LANL, has been taken into account. While the "local" use case is not the primary focus of the ARC file format, we feel it should be seriously taken into account. Ongoing efforts at the Los Alamos National Laboratory (LANL), conducted in the context of an NDIIP project granted by the Library of Congress, reveal the attractiveness of the ARC file format for a "local" use case. We feel that, in light of long term preservation, it can only be beneficial that both crawled and local content could be stored in the same file format.

In the remainder of this document, we provide:

- In Section 2: General comments to the proposed ARC file format revision
- In Section 3: An alternative proposal
- In Section 4: Examples illustrating the alternative proposal.

2. General comments to the proposed ARC file format 3.0 revision

2.1 ari URI

We perceive the following issues with the proposed ari URI approach:

- The identifier is not opaque, but rather of a semantic nature. Elaborate discussions over the past years have led to a significant level of agreement that semantically loaded identifiers should be avoided, whenever possible.
- In order for the identifier to be compliant with the URI syntax (RFC 2396, http://www.faqs.org/rfcs/rfc2396.html), hex-encoding will have to be applied. Given the structure of the identifier and the nature of its components, this seems problematic.
- Reference is made to the Pandora identifiers. Recent registration of the identifier namespace of the National Library of Australia (nla) under the info URI scheme (Internet Draft http://www.ietf.org/internet-drafts/draft-vandesompel-info-uri-02.txt ; Registry http://info-uri.info) revealed problems with the Pandora identifier. As a result, registration of Pandora identifiers under the info:nla/ Scheme was postponed. Further details can be obtained from Debbie Campbell <dcampbel@nla.gov.au>.
- The "embedding" of multiple ari URIs gives a quite inelegant and problematic impression.
- As the ARC file format is intended to be used for long-term use, and the identification of ARC records is of crucial importance, it must be expected that the URI scheme used to identify ARC records will have to be documented in an RFC. Based on personal experience with the info URI scheme, the authors would like to point out that the effort of getting an RFC for a URI scheme accepted is far from trivial, and that serious technical challenges can be expected especially because of the very nature and syntax of the proposed ari scheme.

2.2 ARC Record Serial Number approach

The proposed serial number approach is not waterproof: duplicates can still be created in parallel processes. While the possibility of this happening in the current technological environment may seem low, they can only increase as system performance further improves with time. As waterproof alternatives for the desired functionality exist, it seems unnecessary to take the risk involved in using the proposed Serial Number.

2.3 Usage of Content-type

Three significant problems are perceived in this realm:

- Content-type in protocol responses: The Content-type of the ARC record that contains the response to a protocol request indicates the mime type of the resource that is returned in the protocol response, not the mime type of the protocol response. This problem is present in the existing ARC file format, and is prolonged in the proposed revision. It is truly problematic in light of the long-term use of the ARC file format with protocols (existing and future) other than HTTP that come (and will come) with a variety of response headers. Indeed, the current and proposed ARC file formats do not have the capability to indicate the mime type of a protocol response header. It seems that a future ARC file format should strive to facilitate the unambiguous expression of both the mime type of

the protocol response header and that of the resource provided in the protocol response.

- Proposed new uses of Content-type: While the existing use of the Content-type is already problematic, as described above, it is worsened in the proposal through the proposed use of the mime type of the protocol response header (message/http;msgtype=response) only in those cases that duplicate reduction is intended (Section 6 of the proposal). The result is that, in the proposed format, in some cases the mime type indicates that of the protocol response, in other cases that of the resource returned in the protocol response. This introduced ambiguity, along with the asymmetry of the proposed approach for protocol responses and that for protocol requests strikes us as truly problematic.
- The proposal includes the introduction of an invented, optional qualifier - arcrecordtype – to the mime type (Section 4.1), as a means to identify the type of an ARC record (metadata, duplicate, and transform). This proposal seriously overloads the use of mime types, and the chosen mime type will only be understood by ARC processors. We argue that a special field should be used to convey ARC records types, and that mime type should not be overloaded with ARC-specific semantics.

2.4 Gzip of ARC files

Informative only:

In the repository infrastructure of LANL, so-called XMLtapes play an important role. XMLtapes are files similar to ARC files but contain the concatenation of many individual XML records, each of which acts as a proxy (contains pointers) to real bitstreams that are stored in ARC files. Initial XMLtape implementations were block-zipped, and indexed as such. Zipping resulted in a 90% reduction of the file size. However, comments from several parties, including the Library of Congress NDIIP project team, indicated a concern with zipping the XMLtapes for the purpose of long-term preservation. As a result, efforts are ongoing at LANL aimed at creating tools to deal with the XMLtapes in their native, non-zipped format. Existing XMLtapes, currently used in the production environment, will be unzipped for future use.

Even if the Gzip format is used, we argue that the size of compressed records should be stored in separate index files (cf. Section 8.2 of the proposal: Tom Emerson's proposal), because the offset and size information are related to the indexing of ARC files, not to the ARC files themselves.

2.5 Use of RDF

Informative only:

We would like to express our concern regarding the proposed introduction of RDF in a file format that is currently inherently simple, and that can be processed without the need

for elaborate tools. It seems that the required metadata could easily be conveyed in a text-only manner.

3. Proposed alternative

In the proposed approach, we introduce:

- A refined model to investigate and tackle the problem,
- An alternative mechanism to identify ARC records,
- A separate element to typify the nature of ARC records,
- An unambiguous and symmetric use of mime type shared by all use cases (regular crawl, duplicate reduction, resource transform, local use).

Figure 1 shows a refined model to deal with ARC records. The model is introduced as a means to deal in a consistent manner with various use cases of ARC files and ARC records. It introduces the notion of a transaction of a resource: a resource identified by URI can be processed in a transaction. A sample transaction is a crawl of the resource. Another transaction is the transformation of the resource. In the proposed model, each transaction is accorded its own ARC record identifier. Multiple ARC records may be related to a single transaction (e.g. 3 ARC records can be involved in a crawl transaction: a protocol request ARC record, a protocol response ARC record, a metadata ARC record), and – in the proposed model - all these ARC records share the same ARC record identifier. It is good to think of the ARC record identifier as the unique identifier of the transaction in which the resource is involved.

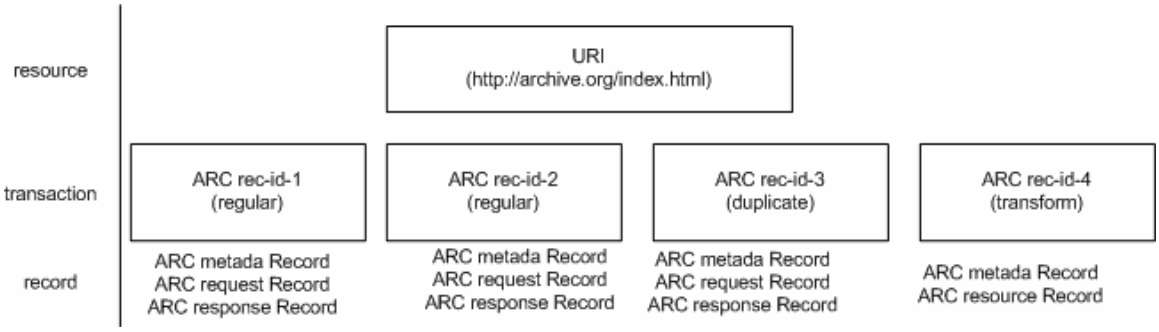

Figure 1. A model for ARC Records

Table 1 shows the different use cases of ARC files, and the nature of the ARC records involved in each use case. As can be see, the Transform use case is very similar to the Local use case. In our proposal, all use cases are handled in a unique, consistent manner.

	Metadata	Protocol Request	Protocol Response (header + resource)	Protocol Response (header only)	Resource only
Regular Crawl	>= 0	1	1	0	0
Duplicate	.>= 1	1	0	1	0

Transform	>= 1	0	0	0	1
Local	>= 0	0	0	0	1

Table 1: Use cases and the nature of ARC records involved

The proposed ARC Record Metadata Line looks as follows:

<URI> <IP-address> <Archive-date> <Content-type> <ARC rec-id> <ARC rec-type> <Checksum> <Archive-length> <nl>

Hereby:

- <URI> : replaces <URL> as a more general designator of the identifier of a resource. The <URI> field contains the URI of the resource for all the ARC records that pertain to the current transaction of the resource.

- <Content-type> : Contains the mime type of the content of the ARC record. This field can take one of the following values:
  - In case the ARC record contains a protocol response (Regular; Duplicate): the mimetype of the protocol response
  - In case the ARC record contains a resource only (Transform; Local): the mime type of the resource
  - In case the ARC record contains a protocol request: the mime type of the protocol request
  - In case the ARC record contains metadata: the mime type of the contained metadata

- <ARC rec-id> : A globally unique identifier computed when ARC records are created, and expressed as a URI. All ARC records related to the transaction of a specific resource share the same value for <ARC rec-id>. For example:
  - In the Regular use case of Table 1, the optional metadata ARC record(s), the protocol request ARC record and the protocol response ARC record share the same value for <ARC rec-id>
  - In the Transform use case of Table 1, the - one or more - metadata ARC records, and the ARC record containing the transformed resource share the same value for <ARC rec-id>

  A proposed implementation for <ARC rec-id> is:
  - Use of the UUID algorithm as described in http://www.ietf.org/internet-drafts/draft-mealling-uuid-urn-03.txt. A typical UUID looks like **00002358-d05c-11d8-85e1-d1cbfd475562 .** Software to create UUIDs exists. Amongst others LANL has created such software, and is willing to share it with the ARC community.
  - Registration and use of a namespace under the info URI scheme, e.g. info:arc/
  - Providing the <ARC rec-id> as **info:arc/arcid/00002358-d05c-11d8-85e1-d1cbfd475562 or info:arc/00002358-d05c-11d8-85e1-d1cbfd475562**

- <ARC rec-type>: A field used to type the ARC record. Values for this field are:
  - In case the ARC record contains a protocol response (Regular; Duplicate): the value **response**
  - In case the ARC record contains a resource only (Transformed; Local): the value **resource**
  - In case the ARC record contains a protocol request: the value **request**
  - In case the ARC record contains metadata: the value **metadata.** Alternatively, the nature of the metadata could be indicated too, e.g. **metadata.duplicate** , **metadata.transform.** However, the type of metadata could be expressed within the metadata record too.

- <Checksum>: Proposed implementation is to register and use a namespace under the info URI scheme, e.g. **info:arc/digest/sha1:…..** or **info:digest/sha1:……**

In Table 2, we illustrate how the fields of the ARC Record Metadata Line are used in the aforementioned use cases. In the Section 4, we provide examples.

	<URI>	<Content-type>	<ARC rec-id>	<ARC rec-type>
Regular				
	url-1	message/http;msgtype=request	uuid-1	request
	url-1	message/http;msgtype=response	uuid-1	response
	url-1	e.g. application/rdf+xml	uuid-1	metadata
Duplicate				
	url-1	message/http;msgtype=request	uuid-2	request
	url-1	message/http;msgtype=response	uuid-2	response
	url-1	e.g. application/rdf+xml	uuid-2	metadata
Transformed				
	url-1	e.g. text/sgml	uuid-3	resource
	url-1	e.g. application/rdf+xml	uuid-3	metadata
Local				
	uri-2	e.g. application/pdf	uuid-4	resource
	uri-2	e.g. application/rdf+xml	uuid-4	metadata

Table 2: Usage of the ARC metadata line fields in different use cases

4. Examples

The following examples use the proposed ARC Record metadata Line.

4.1 Regular use case

ARC record containing the protocol request:

```
http://www.archive.org/index.html 201.201.201.111
20040724100450 message/http;msgtype=request
```

```
info:arc/arcid/00002358-d05c-11d8-85e1-d1cbfd475562 request
info:arc/digest/sha1:abcd 46
GET /index.html   HTTP/1.1
Host: www.archive.org
```

ARC record containing the protocol response:

```
http://www.archive.org/index.html 201.201.201.111
20040724100457 message/http;msgtype=response
info:arc/arcid/00002358-d05c-11d8-85e1-d1cbfd475562
response info:arc/digest/sha1:xxxx 2966
HTTP/1.1 200 OK
Content-type: text/html
Etag: "45184-1f91-41756e9c"
Date: Tue, 03 Aug 2004 01:26:42 GMT
Server: Apache/2.0.47 (Unix) mod_ssl/2.0.47 OpenSSL/0.9.7c
PHP/4.3.4
...
```

4.2 Duplicate reduction

The previously crawled resource is the one listed in the first example. In the ARC records related to the duplicate, the original URI is obviously maintained. The metadata ARC record contains a pointer to the previously crawled resource. The pointer is the <ARC rec-id> of that previously crawled resource: info:arc/arcid/00002358-d05c-11d8-85e1-d1cbfd475562]

ARC record containing the protocol request:

```
http://www.archive.org/index.html 201.201.201.111
20040811090322 message/http;msgtype=request
info:arc/arcid/11112799-e05d-11d8-85e1-a2kdue484632 request
info:arc/digest/sha1:efgh 46
GET /index.html   HTTP/1.1
Host: www.archive.org
If-Modified-Since: Sat, 29 Oct 2006 19:43:31 GMT
```

ARC record containing the protocol response:

```
http://www.archive.org/index.html 201.201.201.111
20040811090322 message/http;msgtype=response
info:arc/arcid/11112799-e05d-11d8-85e1-a2kdue484632
response info:arc/digest/sha1:dfef 66
HTTP/1.1 304 Not Modified
Etag: "45184-1f91-41756e9c"
Date: Mon, 01 Nov 2004 17:57:24 GMT
```

ARC record with metadata for duplicate indication:

```
http://www.archive.org/index.html 201.201.201.111
20040724100459 application/rdf+xml info:arc/arcid/11112799-
e05d-11d8-85e1-a2kdue484632 metadata
info:arc/digest/sha1:zsef 123
<?xml version='1.0'?>
<rdf:RDF
xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:arc='http://www.archive.org/arc/1.1'>
...
<dcterms:isVersionOf>info:arc/arcid/00002358-d05c-11d8-
85e1-d1cbfd475562
</dcterms:isVersionOf>
...
</rdf:RDF>
```

4.3 Resource transformation use case

The original resource is the one listed in the first example. In the ARC records related to the transform, the original URI is maintained. The metadata record contains a pointer to the original resource. The pointer is the <ARC rec-id> of previously existing resource: info:arc/arcid/00002358-d05c-11d8-85e1-d1cbfd475562]

ARC record containing the transformed resource:

```
http://www.archive.org/index.html 192.168.34.55
20051222090322 text/sgml info:arc/arcid/33332799-e44f-21a4-
88b3-b3kdeu864098 resource info:arc/digest/sha1:efgh 3452
...
```

ARC record with metadata for indication of transformation:

```
http://www.archive.org/index.html 192.168.34.55
20051222090324 application/rdf info:arc/arcid/33332799-
e44f-21a4-88b3-b3kdeu864098 metadata
info:arc/digest/sha1:zsef 165
<?xml version='1.0'?>
<rdf:RDF
xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:arc='http://www.archive.org/arc/1.1'>
...
```

```
<dcterms:replaces>info:arc/arcid/00002358-d05c-11d8-85e1-
d1cbfd475562
</dcterms:replaces>
...
</rdf:RDF>
```

4.4 Local use case

ARC record containing the resource:

```
info:doi/10.1045/february2004-bekaert 0.0.0.0
20051222090322 text/html info:arc/arcid/44444799-e44f-21a4-
88b3-b3kdeu864098 resource info:arc/digest/sha1:kdldo 2376
...
```